UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 14-10
Statistics and Econometrics Series (06)
May 2014

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

INDEPENDENT COMPONENTS TECHNIQUES BASED ON KURTOSIS FOR FUNCTIONAL DATA ANALYSIS

Daniel Peña[(1)], Javier Prieto[(2)] and Carolina Rendón[(3)]

Abstract

The motivation for this paper arises from an article written by Peña et al. [40] in 2010, where they propose the eigenvectors associated with the extreme values of a kurtosis matrix as interesting directions to reveal the possible cluster structure of a dataset.

In recent years many research papers have proposed generalizations of multivariate techniques to the functional data case. In this paper we introduce an extension of the multivariate kurtosis for functional data, and we analyze some of its properties. In particular, we explore if our proposal preserves some of the properties of the kurtosis procedures applied to the multivariate case, regarding the identification of outliers and cluster structures. This analysis is conducted considering both theoretical and experimental properties of our proposal.

Keywords: Functional Data Analysis, Functional Kurtosis, Cluster Analysis, Kurtosis Operator.

(1) Peña, Daniel, Department of Statistics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain. e-mail: dpena@est-econ.uc3m.es.
(2) Prieto, Javier, Department of Statistics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain. e-mail: fjp@est-econ.uc3m.es.
(3) Rendón, Carolina, Department of Statistics, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés (Madrid), Spain. e-mail: jrendon@est-econ.uc3m.es.

Independent Components Techniques Based On Kurtosis For Functional Data Analysis

Daniel Peña, Javier Prieto and Carolina Rendón

May 9, 2014

Abstract

The motivation for this paper arises from an article written by Peña et al. [40] in 2010, where they propose using the eigenvectors associated with the extreme values of a kurtosis matrix as interesting directions to reveal the possible cluster structure of a dataset.
In recent years many research papers have proposed generalizations of multivariate techniques to the functional data case. In this paper we introduce an extension of the multivariate kurtosis for functional data, and we analyze some of its properties. In particular, we explore if our proposal preserves some of the properties of the kurtosis procedures applied to the multivariate case, regarding the identification of outliers and cluster structures.
This analysis is conducted from both a theoretical and an experimental point of view, to determine the optimality separation properties of the method for mixtures of gaussian processes, and to evaluate its practical performance on simulated data.

Key words: Functional Data Analysis, Functional Kurtosis, Cluster Analysis, Kurtosis Operator.

---

1 Introduction

Different techniques in multivariate analysis have been designed to reduce the dimensionality of the data and to help derive a simple description of a dataset. Most of them proceed by defining a small number of new variables that summarize the information contained in the original variables. One of the most popular techniques for dimensionality reduction is Principal Component Analysis.
Another problem of interest is the identification of relevant structures in the data: for example, identifying clusters. In these situations, when we assume heterogeneity in the data, the use of Principal Components may not give good results in practice, see [37], and numerous alternatives have been proposed in the literature. A general approach to address the problem of identifying heterogeneity within the framework of dimension reduction is the use of Independent Component Analysis techniques [21], [50]. In these cases, linear combinations of the variables with properties of interest are sought, such as for example, those directions corresponding to projections with the largest possible independence between them.
A particular case is proposed by Peña and Prieto (2001) [38]. They describe a procedure to identify clusters in multivariate data using information obtained from the univariate projections of the sample data on the directions that minimize and maximize the kurtosis coefficient of the projected data. Under certain conditions, these directions have optimal properties to visualize the different clusters that may be present in data. Alternatively, related directions can be obtained from a matrix representation of the kurtosis, with certain implementation advantages. Peña et al. (2010) [40], propose the eigenvectors associated with the extreme values of a kurtosis matrix as interesting directions to reveal the possible cluster structure of a dataset.
Beyond the study of multivariate data, an area of recent interest has been the development of new statistics for Functional Data Analysis (FDA) techniques. In this case, the data, instead of being a vector set, as in classical multivariate analysis, is a set of functions. The purpose of the analysis is to make use of any time (or other independent variable) dependency structure induced by the functions generating the data to obtain a better measure of those aspects of interest in these data. Functional data appears in many fields of application of statistics such as health sciences, economics, environment, among others.
Well-known references in the field of FDA are the books written by Ramsay and Silverman (1997) [45] and Ferraty and Vieu (2006) [15]. In 2005, Ramsay and Silverman [46] wrote a second book of a more applied character in which solutions to the problems associated to concrete datasets are studied. The same authors presented a considerable number of applications in another book [44]. A recently published reference by Ramsay and Hooker includes many Functional Data Analysis applications and algorithmic implementations in R and MATLAB [43].
A random variable X is called a functional variable if it takes values in an infinite dimensional space of functions satisfying some appropriate conditions, and known as a functional space. An observation x of X is called a functional variable. A functional

dataset $x_1, ..., x_n$ is the observation of n functional variables $X_1, ..., X_n$ identically distributed as X.
Due to the high-dimensionality of these functional data, they are usually approximated through their finite expansion in some appropriate (usually orthonormal) basis. A finite number K of terms in the expansion are chosen to represent data in a finite subspace, transforming the infinite dimension problem into a multidimensional problem. The choice of both the parameter K and the most appropriate basis for the observed data is a basic one in functional data analysis, and up to now there is no universal rule providing an optimal selection. The value K acts as a smoothing parameter for the functional data. If K is small we have a very tractable model but possibly relevant information is lost. If K is big, the data are represented with high precision but the computational dimension problem becomes important.
A base is a set of known functions $\{\phi_k\}(k \in \mathbb{N})$ such that any function can be approximated as well as desired, using a linear combination of K of them with K large enough. Thus, a functional observation can be approximated as $x(t) \approx \sum_{k=1}^{K} c_k \phi_k(t)$, where $\{\phi_k\}_{k=1}^{K}$ is a set of base functions and $\{c_k\}_{k=1}^{K}$ is the corresponding set of coefficients.
The most usual bases in functional data analysis are the Fourier basis, B-Splines bases, Wavelets bases, exponential functions, polynomial bases, among others, see Ramsay y Silverman (1997) [45].
Functional Data Analysis (FDA) comprises all the statistical techniques developed for the analysis of curves or surfaces that vary in time. Initially, the research in this area was intended to be an almost direct extension of the techniques of classical multivariate analysis. However, the special structure associated to the functional data implies the need for adapted techniques, and motivates the development of new methodologies and procedures.
As it was mentioned above, Ramsay and Silverman [45] developed an adaptation of Principal Component Analysis to the functional case, the Functional Principal Component Analysis (FPCA) technique. This dimension reduction technique summarizes the information available in the data by identifying a finite set of scalar variables obtained as generalized linear combinations of the curves with maximum variance. However, the technique has well-known shortcomings, such as a high sensitivity to the occurrence of outliers. Also, the summarizing combinations can be difficult to interpret and do not always provide a completely understandable presentation of the structure of the variability in the observed data.
In this paper we will consider extensions of a class of methods that generalize the ideas behind PCA in the multivariate case: independent component analysis. We will apply a version of these methods based on the kurtosis to the unsupervised classification of functional data. The goal of unsupervised classification, given a random sample generated from a mixture of unknown distributions, is to group the sample elements while trying to achieve maximum homogeneity in each group and the largest difference between the groups. For the functional case, if we have a sample obtained from a mixture of several populations, the problem can be enunciated as dividing the functions into groups representing each population.

We propose to introduce a kurtosis operator defined as an extension of the multivariate matrix kurtosis operators. We will interpret and analyze this kurtosis operator for functional data, and we will identify some possible applications for it. In particular, we wish to determine if, regarding the identification of outliers and cluster structures, our proposal can achieve similar to those obtained for the multivariate case mentioned above. Additionally, we want to compare our proposed method with Functional Principal Component Analysis.

Classification for functional data has been recently considered by several authors. One of the early references on the subject was that of Hastie et al. (1995) [22]. They adapt the general ideas for functional discriminant analysis, based on a penalized method for regularization. This setting allows them to cast the classification problem as a regression problem via optimal scoring. This facilitates the use of any penalized regression technique in the functional classification setting.

In the context of unsupervised classification, K-Means was one of the first methods to be adapted to the functional case. Various implementations and variations have emerged, among them those by Abraham et al.(2003) [1], where they propose a clustering method consisting of the fitting the functional data using B-splines and partitioning the estimated model coefficients using a K-means algorithm. Biau et al. (2005) [5] obtained results on K-means in infinite dimensional Hilbert spaces, where they propose using a nonparametric method and describe the problem of functional classification as a generalization of the classification problem of the elements of $\Re^d$ to the random variables X_i, taking values in a separable infinite-dimensional Hilbert space.

James and Sugar (2003) [24] develop a flexible model-based procedure for clustering functional data. The technique can be applied to all types of curve data but is particularly useful when individuals are observed at a sparse set of time points. Also they extend the model to handle multiple functional and finite dimensional covariates.

Serban and Wasserman (2005) [48], propose a technique for nonparametrically estimating and clustering a large number of curves called *CATS: Clustering After Transformation and Smoothing*. In this method they estimate the error due to the fact that we are clustering the estimated curves rather than the true curves. CATS is quite general, but they describe and analyze the method mostly in the context of microarray experiments.

In the framework of supervised classification some extensions have also been made to the functional case. It is worth mentioning the study of Ferraty and Vieu (2003) [14], where they propose a nonparametric supervised classification model by introducing a consistent kernel estimator, but applied to a sample of curves. López-Pintado and Romo (2006) [27] consider the role of continuity of data and propose robust procedures based on the concept of depth for the supervised classification of curves. Recently, Baíllo et al. (2011) [3] shown that an optimal classification rule can be explicitly obtained for a class of Gaussian processes with "triangular" covariance functions.

Moreover, Hall et al. (2001) [18], employ a functional data analytic method for dimension reduction based on Principal Component Analysis (PCA) and perform Quadratic Discriminant Analysis (QDA) in the reduced space. Ramsay and Silverman analyze similar techniques, see [44] and [45]. Yao et al.(2004) [52] propose a nonparametric method to

perform functional principal component analysis for the case of sparse longitudinal data. Song et al.(2007) [49] describe a method based on functional data analysis to cluster time-dependent gene expression profiles. Chiou and Li [7] introduced a functional clustering (FC) method for longitudinal data, called *k-centres FC*, and showed that, under the identifiability conditions they derived, the k-centres FC method can greatly improve cluster quality as compared to conventional clustering algorithms. Furthermore, by exploring the mean and covariance functions of each cluster, the k-centres FC method provides an additional insight into cluster structures which facilitates functional cluster analysis.
A significant number of papers on the related topic of outlier detection for functional data have also been published. Ramsay and Silverman (1997) [45] have developed the Principal Component Analysis for functional data, to identify atypical isolates, just as in the multivariate case. But there is no assurance that this method works when there are groups of atypical observations due to the problem of masking, as in the multivariate case. Moreover, Febrero et al. (2007,2008) [12], [13], Martínez et al.(2011) [30], Díaz et al.(2012) [11] and Jacques and Preda (2012) [23], among others, have made some advances in outlier detection for functional data with applications in different areas.
For the multivariate analysis case, the kurtosis has been used as a way to treat the heterogeneity, or to detect the presence of outliers. Peña et al. (2010) [40] propose the eigenvectors associated with the extreme values of a kurtosis matrix as interesting directions to reveal the possible cluster structure of a dataset. In this paper we adapt this approach based on the kurtosis for the identification of outliers and cluster structures for functional data.
The paper is structured as follows: Section 2 provides a general description of the proposed method. Some of its most relevant theoretical properties are also analyzed in Section 3. In Section 4 the results of some computational experiments to compare the performance of the proposed method with FPCA are presented, as well as other results of the application of the proposed operator on some real-life datasets. We finish with some remarks and conclusions in Section 5.

2 Description of the Kurtosis operator

2.1 Interpretations of the kurtosis for univariate and multivariate data

In symmetrical univariate models, the kurtosis is a measure of the peakedness of the probability distribution of a real-valued random variable. Its value also reflects the presence of heavy tails or bimodality in the data. These properties allow the use of the kurtosis for the identification of the possible cluster structure and the existence of outliers in a data set.
While the definition of the kurtosis for the univariate case is well established from the work of Pearson (1905) [36] or Darlington (1970) [9], there is no single way to define the kurtosis in the multivariate case. From the different alternative proposals we mention the works of Móri et al. (1993) [34] and Peña (2002) [37], as the ones providing the most direct reference for our extension to the functional case.

In particular, in [34] a kurtosis matrix for a multivariate random variable X is defined as

$$K = ZZ^T ZZ^T, \quad Z = \Sigma^{-1/2}(X - E[X]),$$

where $\Sigma = \mathrm{Var}(X)$. We adapt this definition to the case in which we have a sample of functional data observations.

2.2 A kurtosis operator for Functional Data

Let $x_1(t), ..., x_n(t)$ be a set of functional data observations in a Banach space with inner product $< \cdot, \cdot >$.
We define a kurtosis kernel for this data as

$$k(s,t) = \frac{1}{n} \sum_i < x_i, x_i > x_i(s) x_i(t).$$

And the associated kurtosis operator $K(z)$ as

$$K(z) = \frac{1}{n} \sum_i < x_i, x_i >< x_i, z > x_i. \tag{1}$$

This operator is linear and its eigenfunctions $\xi(t)$ satisfy

$$\int k(s,t)\xi(t)dt = \lambda \xi(t),$$

for an infinite number of (real) eigenvalues λ. Note that this operator is also positive definite, as

$$< z, K(z) >= \frac{1}{n} \sum_i < x_i, x_i > (< x_i, z >)^2 \geq 0 \quad \forall z.$$

2.3 Implementation of the proposed kurtosis operator

In this Section we describe in detail how we conduct in practice the implementation of the calculations required to obtain the values corresponding to the application of the proposed kurtosis operator to a sample of functional data. At the same time, we provide a link between the application of the kurtosis operator to functional data and the use of the kurtosis matrix for multivariate data.
We assume we are given a sample of multivariate observations, generated from a functional data model. These data have the form

$$x_i(t_j), \quad i = 1, ..., n, \quad t_j \in [0, T], \quad j = 0, ..., p$$

The application of the proposed operator is carried out in a series of steps, that are enumerated and described below.

1. *Representation.* As we mentioned in the introduction, we wish to take advantage of the structure of the data as functional objects, to improve on any results that could be obtained from any direct treatment of the data as multivariate objects. For example, we may wish to conduct some exploratory analysis to identify the main characteristics of the data or prepare these data for later treatments [45]. Alternatively, we may wish to perform a cluster analysis, which is our main motivation for this proposal. Our first step will be to find a reasonable functional representation for our data.

   To obtain this representation, we start by selecting a functional basis. Let $\phi_k(t)$ for $t \in [0,T]$ and $k = 1, ..., m$ denote a truncated basis. Any function can be approximated to arbitrary precision as a linear combination of the functions in this basis as long as m is chosen to be large enough. We select a value for m providing a reasonable balance between precision and complexity.

   We obtain values for a set of coefficients c_{ik} (using regularized least-squares, or some other related method) such that

$$x_i(t) \approx \hat{x}_i(t) = \sum_k c_{ik}\phi_k(t).$$

   In matrix form we can write $\hat{x} = C\phi$, for $\hat{x}$ and ϕ vectors of functional values and $C \in \mathbb{R}^{n\times m}$. To simplify this presentation, we assume that the number of observations for each function is the same, or at least that we work with the same numbers of observations from each function in the smoothed data.

2. *Centering the functional data.* We subtract the mean from the data,

$$\begin{aligned}
\bar{x}(t) &= \frac{1}{n}\sum_i \hat{x}_i(t) = \frac{1}{n}\sum_{ik} c_{ik}\phi_k(t) = \sum_k \left(\frac{1}{n}\sum_i c_{ik}\right)\phi_k(t) \\
\tilde{x}_i(t) &= \hat{x}_i(t) - \bar{x}(t) = \sum_k \left(c_{ik} - \frac{1}{n}\sum_l c_{lk}\right)\phi_k(t) = \sum_k \tilde{c}_{ik}\phi_k(t).
\end{aligned}$$

   This operation can be written in matrix form as $\tilde{x} = (I - \frac{1}{n}ee^T)C\phi = \tilde{C}\phi$

3. *Transforming the data.* Our next step will remove from the data any variability information that might be present. Any relevant pattern directly associated to this variability can be analyzed using Principal Component techniques, for example. Our goal is to go beyond these patterns to reveal additional structure in the data, such as outliers or clusters, that might be hidden in the variability information.

   We identify a linear transformation with kernel $l(s,t)$ that will provide us with the desired transformed observations $y_i(t)$,

$$y_i(t) = \int l(s,t)\tilde{x}_i(s)ds.$$

In particular, we wish the transformed functions y_i to have a unitary covariance operator.

Let $\bar{\phi}_i(t)$ denote the basis functions transformed using the operator l,

$$\bar{\phi}_i(t) = \int l(s,t)\phi_i(s)ds.$$

From these functions we generate a new basis $\tilde{\phi}_i(t)$ projecting the transformed functions onto span(ϕ_i). We write $\tilde{\phi} = A\phi$ for some matrix A associated with $l(s,t)$, and we assume this matrix to be invertible.

We will have $y = \tilde{C}\tilde{\phi} = \tilde{C}A\phi = \hat{C}\phi$. To identify the form of A that ensures the desired properties for $y(t)$, let W be defined as $W_{ij} =< \phi_i, \phi_j >$. Our desired matrix is

$$A = \sqrt{n-1}(\tilde{C}^T\tilde{C})^{-1/2}W^{-1/2},$$

ensuring that $E[y] = 0$ and for any $z = \gamma^T\phi$,

$$\begin{aligned}\frac{1}{n-1}\sum_i < y_i, z > y_i &= \frac{1}{n-1}\sum_i < \sum_{kl}\tilde{c}_{ik}a_{kl}\phi_l, \sum_r \gamma_r\phi_r > \sum_{st}\tilde{c}_{is}a_{st}\phi_t \\ &= \frac{1}{n-1}(\tilde{C}AW\gamma)^T\tilde{C}A\phi \\ &= \gamma^T\phi = z.\end{aligned}$$

4. *Kurtosis operator.* For the kurtosis operator defined as in (1), and an arbitrary function $z = \gamma^T\phi$ we have that

$$K(z) = \frac{1}{n}\sum_i\sum_{kl}\hat{c}_{ik}\hat{c}_{il} < \phi_k, \phi_l > \sum_{rs}\hat{c}_{ir}\gamma_s < \phi_r, \phi_s > \sum_t \hat{c}_{it}\phi_t,$$

or in matrix form,

$$K = \frac{1}{n}\left(\hat{C}^T D\hat{C}W\gamma\right)^T\phi,$$

where D is a diagonal matrix with entries $D_{ii} = \|y_i\|^2$, that is, $D = \text{diag}(\hat{C}W\hat{C}^T)$.

The eigenfunctions and eigenvalues of this operator can be characterized as $K(z(t)) = \lambda z(t)$, or in our equivalent matrix form

$$\frac{1}{n-1}\hat{C}^T D\hat{C}W\gamma = \lambda\gamma.$$

Numerically, it may be more efficient to obtain the eigenvalues from the symmetric matrix

$$K_f \equiv \frac{1}{n}W^{1/2}\hat{C}^T D\hat{C}W^{1/2}, \tag{2}$$

and the eigenfunctions will be obtained from the eigenvectors of this matrix, $\hat{\gamma}$, as

$$K_f\hat{\gamma} = \frac{1}{n}W^{1/2}\hat{C}^T D\hat{C}W^{1/2}\hat{\gamma} = \lambda\hat{\gamma}$$

using the transformation $\gamma = W^{-1/2}\hat{\gamma}$.

This representation of the (approximate) eigenvectors and eigenvalues of the kurtosis operator allows for an interesting comparison with the direct application of the kurtosis matrix proposed by Móri et al. [34] to the original multivariate data, $x_i(t_j)$ (or to the smoothed data $\hat{x}_i(t_j)$).
To study this case, let $X \in \mathbb{R}^{n\times p}$ denote the matrix of multivariate observations. We introduce

$$\tilde{X} = \left(I - \tfrac{1}{n} ee^T\right) X, \qquad Z = \tilde{X}\left(\tilde{X}^T\tilde{X}\right)^{-1/2}.$$

The multivariate kurtosis matrix is defined as

$$K_m = \frac{1}{n}\sum_{i=1}^{n}(z_i^T z_i) z_i z_i^T = \frac{1}{n} Z^T D_Z Z, \tag{3}$$

where $D_Z = \mathrm{diag}(z_i^T z_i) = \|z_i\|^2$.
If we compare (3) and (2), we may conclude that the application of the functional kurtosis procedure is closely related to using the Móri multivariate kurtosis matrix K_m, computed from the values of the coefficients in the basis expansion of the functional data. In particular, both procedures coincide when $W = I$, that is, when we represent our data using an orthogonal basis.

3 Some theoretical properties of the functional kurtosis operator

In this section we analyze the properties of the proposed kurtosis operator with respect to the optimal classification of observations from mixtures of distributions. The general problem for any distribution in the observations is too complex for us to be able to get any meaningful results in that case, and we will consider a simplified case. We will make use of an analogy with the multivariate case, where the simplest situation is that corresponding to a mixture of two normal distributions with the same covariance matrix, that is, the reference case giving rise to the Fisher discriminant function.
In the functional case we will study the equivalent to that case: the situation when the (functional) data have been obtained from a mixture of two gaussian processes with the same covariance operator.

Optimal classification rules for a mixture of gaussian processes

Consider a model in which you sample n observations from a mixture of two gaussian processes. To be more precise, consider a gaussian process on the interval $[0, 1]$ with mean function $m_1(t)$ and covariance function $k(s, t)$, and a second gaussian process on the same interval with mean function $m_2(t)$ and the same covariance function $k(s, t)$, and generate n observations by selecting an observation from the first group with probability α and from the second group with probability $1 - \alpha$.

Let ϕ_k denote the set of orthogonal eigenfunctions for $k(s,t)$ and $\lambda_k \geq 0$ the corresponding eigenvalues. Then

$$k(s,t) = \sum_{k=0}^{\infty} \lambda_k \phi_k(s)\phi_k(t).$$

Also, let

$$m_1(t) - m_2(t) = \sum_{k=0}^{\infty} \nu_k \phi_k(t).$$

As our first step, we will determine the functional that optimizes a certain separation criterion, that is also optimized by the Fisher discriminant function.
We will work with the functional defined as the ratio of the variability between groups and the variability within the groups, for the functional observations projected onto a given function φ, defined for our data model as

$$\Delta(\varphi) \equiv \frac{BTG(\varphi)}{WTG(\varphi)} = \frac{\alpha(1-\alpha)\langle \varphi, m_1 - m_2 \rangle^2}{\iint \varphi(s)\varphi(t)k(s,t)dsdt}.$$

Lemma 1 *Assume that $\lambda_k > 0$ for all k. The function that maximizes the value of Δ is given by*

$$\varphi(t) = \sum_{k=0}^{\infty} \omega_k \phi_k(t), \qquad \omega_k = C\frac{\nu_k}{\lambda_k},$$

for some constant C.

Proof Given

$$\varphi(t) = \sum_{k=0}^{\infty} \omega_k \phi_k(t),$$

if we rewrite Δ in terms of the eigenfunctions and eigenvalues of k we have

$$\Delta(\omega) = \frac{\alpha(1-\alpha)(\sum_k \nu_k\omega_k)^2}{\sum_k \lambda_k \omega_k^2}.$$

If $\lambda_i > 0$, the first-order optimality conditions are

$$\frac{2\alpha(1-\alpha)(\sum_k \nu_k\omega_k)(\sum_k \lambda_k\omega_k^2)\nu_i - 2\alpha(1-\alpha)(\sum_k \nu_k\omega_k)^2\lambda_i\omega_i}{(\sum_k \lambda_k\omega_k^2)^2} = 0, \tag{4}$$

and

$$\frac{2\alpha(1-\alpha)(\sum_k \nu_k\omega_k)\nu_i}{\sum_k \lambda_k\omega_k^2} = 0.$$

otherwise.
The solutions for these equations are either $\sum_k \nu_k\omega_k = 0$ (the minimizer of the problem) or the one given above. □

Unfortunately, this representation of φ has the undesirable property of not having a bounded norm. As in practice we will work with a finite basis representation of the data,

it would seem interesting to ensure that the norm of the functions we use is bounded, to guarantee good truncation properties. A slight modification of the preceding lemma is given below.

Lemma 2 *The function that solves the problem*

$$\max_{\varphi} \Delta(\varphi) \qquad s.t.\ \|\varphi\| \leq V,$$

is given by

$$\varphi(t) = \sum_{k=0}^{\infty} \omega_k \phi_k(t), \qquad \omega_k = C_1 \frac{\nu_k}{\lambda_k + C_2},$$

for some constants C_1 and $C_2 > 0$.

Proof If we rewrite Δ in terms of the eigenfunctions and eigenvalues of k, as in (4), we have that the problem of interest in this case is

$$\max_{\omega} \frac{\alpha(1-\alpha)(\sum_k \nu_k \omega_k)^2}{\sum_k \lambda_k \omega_k^2} \qquad \text{s.t. } \sum_k \omega_k^2 \leq V^2.$$

Its first-order optimality conditions are

$$\frac{2\alpha(1-\alpha)(\sum_k \nu_k\omega_k)(\sum_k \lambda_k\omega_k^2)\nu_i - 2\alpha(1-\alpha)(\sum_k \nu_k\omega_k)^2\lambda_i\omega_i}{(\sum_k \lambda_k\omega_k^2)^2} + 2\mu\omega_i = 0.$$

The optimal solution for ω_k given above follows from this. □

If there exists a smallest $\tilde{k}$ such that $\lambda_k = 0$ for all $k > \tilde{k}$, the problem has no (bounded) solution on ω_k unless $\nu_k = 0$ for all $k > \tilde{k}$, in which case we again obtain the preceding solution, with $\omega_k = 0$ for $k > \tilde{k}$.

Discriminating properties of some eigenfunctions of a kurtosis operator

In this section we relate the properties of a kurtosis operator to the optimal discriminating properties discussed in the preceding section. Our main goal is the identification of relationships between the eigenfunctions of that kurtosis operator and the preceding optimal classification function. We will follow a procedure very similar to the one presented in Tyler et al. (2009) [50] for the multivariate case.
As in the preceding case, we will work with a process defined as a mixture, with probability α, of two gaussian processes $X_1(t) = m_1(t) + \sum_i \xi_{1i}\phi_i(t)$ and $X_2(t) = m_2(t) + \sum_i \xi_{2i}\phi_i(t)$, having the same covariance operator with kernel function k, which we will write in terms of its eigenvalues and eigenfunctions as

$$k(s,t) = \sum_{k=0}^{\infty} \lambda_k \phi_k(s)\phi_k(t).$$

Let also $E[X] \equiv \mu_X = \alpha m_1 + (1-\alpha)m_2$ and

$$m(t) \equiv m_1(t) - m_2(t) = \sum_{k=0}^{\infty} \nu_k \phi_k(t).$$

We can then write

$$X(t) = \begin{cases} m_1 + \sum_i \xi_{1i}\phi_i & \text{w.p. } \alpha \\ m_2 + \sum_i \xi_{2i}\phi_i & \text{w.p. } 1-\alpha, \end{cases} \tag{5}$$

where ξ_{1i} and ξ_{2i} are independent standard normal random variables.
For any function $u = \sum_i u_i \phi_i$, the covariance operator for X applied to this function will yield

$$S_X(u) = E[\langle X - \mu_X, u\rangle (X - \mu_X)] = \sum_i \left(\lambda_i u_i + \alpha(1-\alpha)\nu_i \sum_j \nu_j u_j \right) \phi_i = \sum_{ij} s_{ij} u_i \phi_j,$$

where

$$s_{ij} = \begin{cases} \lambda_i + \alpha(1-\alpha)\nu_i^2 & \text{si } i = j, \\ \alpha(1-\alpha)\nu_i\nu_j & \text{otro caso.} \end{cases}$$

For any process X, and in particular for the previously defined one, we introduce a kurtosis operator K_X given by

$$K_X(u) = E[\langle X - \mu_X, R_X(X-\mu_X)\rangle \langle X - \mu_X, u\rangle (X - \mu_X)],$$

where R_X is some approximation to the inverse of the covariance operator S_X. Our goal is to show that this kurtosis operator has an eigenfunction closely related to those introduced in Lemmas 1 and 2.
We start with an initial result introducing a characterization of the effect of some (specific) linear transformations of X on the corresponding covariance operators.

Lemma 3 *Assume a linear operator $L(u)$ defined by a kernel of the form $g(s,t) = \sum_i \pi_i \phi_i(s)\phi_i(t)$ for some coefficients π_i, that is,*

$$L(u) = \int g(s,t)u(s)ds = \sum_i \pi_i u_i \phi_i(t). \tag{6}$$

Define $Y = L(X - \mu_X)$. Then, $E[Y] = 0$ and

$$S_Y(u) = \sum_{ij} s_{ij}\pi_i\pi_j u_i \phi_j.$$

Proof Applying this transformation we have

$$\begin{aligned} Y_1(t) &= L(X_1) = \int g(s,t)\left(m_1(s) + \sum_i \xi_{1i}\phi_i(s) - \alpha\mu_1(s) - (1-\alpha)\mu_2(s)\right) ds \\ &= \sum_i \xi_{1i}\pi_i\phi_i(t) + (1-\alpha)\tilde{m}(t) \\ Y_2(t) &= L(X_2) = \int g(s,t)\left(m_2(s) + \sum_i \xi_{2i}\phi_i(s) - \alpha\mu_1(s) - (1-\alpha)\mu_2(s)\right) ds \\ &= \sum_i \xi_{2i}\pi_i\phi_i(t) - \alpha\tilde{m}(t), \end{aligned}$$

where $\tilde{m}(t) = \sum_i \pi_i\nu_i\phi_i(t)$.
Then,

$$\begin{aligned} S_Y(u) &= E[\langle Y, u\rangle Y] = \alpha E\left[\sum_i (\xi_{1i} + (1-\alpha)\nu_i)\,\pi_i u_i \sum_j (\xi_{1j} + (1-\alpha)\nu_j)\,\pi_j\phi_j\right] \\ &\quad + (1-\alpha)E\left[\sum_i (\xi_{2i} - \alpha\nu_i)\,\pi_i u_i \sum_j (\xi_{2j} - \alpha\nu_j)\,\pi_j\phi_j\right] \\ &= \sum_i \pi_i^2\lambda_i u_i\phi_i + \alpha(1-\alpha)\sum_i \left(\sum_j \pi_j\nu_j u_j\right)\pi_i\nu_i\phi_i = \sum_{ij} s_{ij}\pi_i\pi_j u_i\phi_j. \end{aligned}$$

□

An immediate consequence of this result is the following one:

Corollary 1 *If $\pi_i = \lambda_i^{-1/2}$ for all i, then $\tilde{m}$ is an eigenfunction of S_Y. The eigenvalues of S_Y associated to other eigenfunctions are all of them equal to 1.*

Proof From Lemma 3 it holds that

$$S_Y(\tilde{m}) = \sum_i \pi_i\nu_i\phi_i + \alpha(1-\alpha)\sum_j \pi_j^2\nu_j^2 \sum_i \pi_i\nu_i\phi_i = \left(1 + \alpha(1-\alpha)\sum_i \pi_i^2\nu_i^2\right)\tilde{m}.$$

Also, for any function ψ_k such that $\langle\psi_k, \tilde{m}\rangle = \sum_i \langle\psi_k, \phi_i\rangle\pi_i\nu_i = 0$ it holds that $S_Y(\psi_k) = \sum_i \langle\psi_k, \phi_i\rangle\phi_i = \psi_i$. □

To proceed with our proof we need to introduce a condition on the operator R_X introduced to define our kurtosis operator. We will require that the following property holds:

C1. Let X be a stochastic process, and $Y = M(X - \mu_X)$ with $M(u) = \sum_{ij} \varsigma_{ij} u_i\phi_j$. If the operator R can be written as

$$R_Y(u) = \sum_{ij} r_{ij} u_i\phi_j,$$

for some set of values r_{ij}, then it holds that

$$R_X(u) = \sum_{ij} r_{ij}\varsigma_{ik}\varsigma_{jl}u_k\phi_l. \tag{7}$$

Note in particular that for the case when $M = L$ we have that $\varsigma_{ij} = \pi_i$ if $i = j$ and $\varsigma_{ij} = 0$ otherwise, and the condition would require

$$R_X(u) = \sum_{ij} r_{ij}\pi_i\pi_j u_i\phi_j.$$

This condition implies the following result:

Lemma 4 *Let X be a stochastic process that can be written as $X(t) = \mu_X(t)+\sum_i \xi_i\phi_i(t)$, and $Y = L(X - \mu_X)$. If C1 holds then*

$$\langle X - \mu_X, R_X(X - \mu_X)\rangle = \langle Y, R_Y(Y)\rangle. \tag{8}$$

Proof Let $X - \mu_X = \sum_i \xi_i\phi_i$. Then $Y = \sum_i \pi_i\xi_i\phi_i$ and

$$\langle X - \mu_X, R_X(X - \mu_X)\rangle = \sum_{ij} r_{ij}\pi_i\pi_j\xi_i\xi_j = \langle Y, R_Y(Y)\rangle.$$

□

To simplify the derivation of the main results, we define θ as the random variable corresponding (under condition C1) to

$$\theta \equiv \langle X - \mu_X, R_X(X - \mu_X)\rangle = \langle Y, R_Y(Y)\rangle.$$

The following result characterizes the form of the proposed kurtosis operator, for the particular case of the process X introduced at the beginning of the section, and its behavior under transformation L.

Lemma 5 *Let X be the random process introduced in (5) and define*

$$\chi_i = \alpha E[\theta\xi_{1i}^2] + (1 - \alpha)E[\theta\xi_{2i}^2], \quad \chi = E[\theta].$$

Then, for $u = \sum_i u_i\phi_i$,

$$\begin{aligned} K_X(u) &= \sum_{ij} \kappa_{ij}u_i\phi_j \\ \kappa_{ij} &= \begin{cases} \chi_i + \alpha(1 - \alpha)\chi\nu_i^2 & \text{if } i = j, \\ \alpha(1 - \alpha)\chi\nu_i\nu_j & \text{otherwise.} \end{cases} \end{aligned}$$

For $Y = L(X - \mu_X)$ we have

$$K_Y(u) = \sum_{ij} \kappa_{ij}\pi_i\pi_j u_i\phi_j.$$

Proof Our first result is related to the symmetry properties of some moments. For $i = 1, 2$,

$$\begin{aligned} E[\theta\xi_{ij}] &= \alpha \sum_{kl} r_{kl}\pi_k\pi_l E\left[\xi_{1k}\xi_{1l}\xi_{ij}\right] + (1-\alpha)\sum_{kl} r_{kl}\pi_k\pi_l E\left[\xi_{2k}\xi_{2l}\xi_{ij}\right] = 0, \quad \forall j \\ E[\theta\xi_{ij}\xi_{ik}] &= \alpha \sum_{lm} r_{lm}\pi_l\pi_m E\left[\xi_{1l}\xi_{1m}\xi_{ij}\xi_{ik}\right] + (1-\alpha)\sum_{lm} r_{lm}\pi_l\pi_m E\left[\xi_{2l}\xi_{2m}\xi_{ij}\xi_{ik}\right] = 0, \quad \forall j \neq k. \end{aligned}$$

Using these values, the kurtosis operator satisfies

$$\begin{aligned} K_X(u) &= E[\langle X-\mu_X, R_X(X-\mu_X)\rangle\langle X-\mu_X, u\rangle(X-\mu_X)] \\ &= \alpha E\left[\theta\sum_i \left(\xi_{1i}+(1-\alpha)\nu_i\right)u_i \sum_j \left(\xi_{1j}+(1-\alpha)\nu_j\right)\phi_j\right] \\ &\quad + (1-\alpha)E\left[\theta\sum_i \left(\xi_{2i}-\alpha\nu_i\right)u_i \sum_j \left(\xi_{2j}-\alpha\nu_j\right)\phi_j\right] \\ &= \sum_i u_i\chi_i\phi_i + \alpha(1-\alpha)\chi\sum_{ij} u_i\nu_i\nu_j\phi_j. \end{aligned}$$

For the operator applied to Y we have in a similar manner

$$\begin{aligned} K_Y(u) &= E[\langle Y, R_Y(Y)\rangle\langle Y, u\rangle Y] = E[\theta\langle Y, u\rangle Y] \\ &= \alpha E\left[\theta\sum_i \left(\xi_{1i}+(1-\alpha)\nu_i\right)\pi_i u_i \sum_j \left(\xi_{1j}+(1-\alpha)\nu_j\right)\pi_j\phi_j\right] \\ &\quad + (1-\alpha)E\left[\theta\sum_i \left(\xi_{2i}-\alpha\nu_i\right)\pi_i u_i \sum_j \left(\xi_{2j}-\alpha\nu_j\right)\pi_j\phi_j\right] \\ &= \sum_i \pi_i^2 u_i\chi_i\phi_i + \alpha(1-\alpha)\chi\sum_{ij} \pi_i\pi_j u_i\nu_i\nu_j\phi_j. \end{aligned}$$

□

Another result identifies the eigenfunctions corresponding to operators having a certain structure that will be useful for the proof of the main result.

Lemma 6 *Let $\pi_i = \lambda_i^{-1/2}$ and $Y = L(X-\mu_X)$. We consider an operator $T_Y(u) = \sum_{ij} t_{ij}u_i\phi_j$ and any transformation $U(u) = \sum_{ij}\varsigma_{ij}u_i\phi_j$ such that $U(\tilde{m}) = \tilde{m}$ and U is a rotation for any other combination of eigenfunctions.*
If for $\tilde{Y} = U(Y)$ the operator $T_{\tilde{Y}}$ satisfies

$$T_{\tilde{Y}}(u) = \sum_{ijkl} t_{ij}\varsigma_{ik}\varsigma_{jl}u_k\phi_l, \tag{9}$$

then $\tilde{m}$ is an eigenfunction of T_Y.

Proof From the properties of Y in our case, and in particular as S_Y has a distribution that is symmetric with respect to all eigenfunctions except for $\tilde{m}$, its distribution will not change under U, that is, the distribution of $\tilde{Y}$ will coincide with that of Y. In particular, for $\tilde{Y} = U(Y)$ we will have $T_Y = T_{\tilde{Y}}$.
Also, from U being a rotation we have that

$$\sum_k \varsigma_{ik}\varsigma_{jk} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

From $U(\tilde{m}) = \tilde{m}$ we have $\sum_i \varsigma_{ij}\tilde{m}_i = \tilde{m}_j$, but also from the preceding result

$$\sum_i \varsigma_{ji}\tilde{m}_i = \sum_{ik} \varsigma_{ji}\varsigma_{ki}\tilde{m}_k = \tilde{m}_j.$$

It then holds that

$$T_Y(\tilde{m}) = T_{\tilde{Y}}(\tilde{m}) = \sum_{ijkl} t_{ij}\varsigma_{ik}\varsigma_{jl}\tilde{m}_k\phi_l = \sum_{ijl} t_{ij}\varsigma_{jl}\tilde{m}_i\phi_l \sum_{jk}\varsigma_{jk}\sum_i (t_{ij}\tilde{m}_i)\,\phi_k = U(T_Y(\tilde{m})),$$

but this must hold for any U satisfying the indicated properties. Thus, $T_Y(\tilde{m}) = \tilde{m}$. □

We now present the main result for the section.

Theorem 1 *Assume that condition C1 holds. The function $\psi(t)$, defined as*

$$\psi(t) = \sum_i \frac{\nu_i}{\lambda_i}\phi_i(t),$$

is an eigenfunction of the operator $R_X(K_X)$.

Proof

- We start by transforming the random function X into another random function $Y = L(X - \mu_X)$ with a more desirable eigenvalue/eigenfunction structure. To do that, select $\pi_i = \lambda_i^{-1/2}$. From Corollary 1 we know that Y has an eigenfunction given by $\tilde{m}$ and all other eigenvalues are equal to 1.

- The new random function Y is a mixture of two gaussian processes Y_1 and Y_2 and it has a distribution that is invariant to rotations not affecting the first eigenfunction ($\tilde{m}$). By symmetry, any rotation that does not affect the first eigenfunction will not change the covariance operator. As a consequence, all processes obtained by applying any of these rotational transformations to Y will have the same distribution, and in particular their kurtosis operators will be the same.

  We use Lemma 6 to establish that $R_Y(K_Y)$ must have $\tilde{m}$ as an eigenfunction. It is enough to show that $R_{\tilde{Y}}(K_{\tilde{Y}})$ transforms according to (9) for $\tilde{Y} = U(Y)$.

Using Lemma 5 and the definition of R given in condition C1 we have,

$$\begin{aligned} R_Y(K_Y(u)) &= \sum_{ijk} r_{ij}\kappa_{ki}\pi_i\pi_k u_k\phi_j, \\ R_X(K_X(u)) &= \sum_{ijk} r_{ij}\pi_i\pi_j\kappa_{ki}u_k\phi_j. \end{aligned}$$

It may be useful to rewrite $R_Y(K_Y(u)) = \sum_{ij}\tilde{r}_{ij}u_i\phi_j$, where $\tilde{r}_{ij} = \sum_k r_{kj}\kappa_{ik}\pi_i\pi_k$. From condition C1 we have that

$$\begin{aligned} R_{\tilde{Y}}(u) &= \sum_{ijkl} r_{ij}\varsigma_{ik}\varsigma_{jl}u_k\phi_l, \\ K_{\tilde{Y}}(u) &= \sum_{ijkl} \tilde{\kappa}_{ij}\varsigma_{ik}\varsigma_{jl}u_k\phi_l, \end{aligned}$$

where $\tilde{\kappa}_{ij} = \kappa_{ij}\pi_i\pi_j$. Combining these results we obtain

$$\begin{aligned} R_{\tilde{Y}}(K_{\tilde{Y}}(u)) &= \sum_{ijkl} r_{ij}\varsigma_{ik}\varsigma_{jl}\left(\sum_{mno}\tilde{\kappa}_{mn}\varsigma_{mo}\varsigma_{nk}u_o\right)\phi_l = \sum_{ijlmno}\left(\sum_k \varsigma_{ik}\varsigma_{nk}\right) r_{ij}\tilde{\kappa}_{mn}\varsigma_{jl}\varsigma_{mo}u_o\phi_l \\ &= \sum_{ijlmo} r_{ij}\tilde{\kappa}_{mi}\varsigma_{jl}\varsigma_{mo}u_o\phi_l = \sum_{jlmo}\tilde{r}_{mj}\varsigma_{jl}\varsigma_{mo}u_o\phi_l. \end{aligned}$$

This result implies that condition (9) is satisfied. Thus, from Lemma 6 we conclude that $\tilde{m}$ is an eigenfunction of $R_Y(K_Y(u))$.

- Our next step is to identify an eigenfunction of $R_X(K_X)$. Introduce the transformation $v(t) \equiv V(u) \equiv \sum_i \pi_i u_i\phi_i(t)$, and note that it is a linear operator on u. It holds that

$$R_X(K_X(V(u))) = \sum_{ijk} r_{ij}\pi_i\pi_j\kappa_{ki}\pi_k u_k\phi_j.$$

Letting $w = W(u) \equiv R_Y(K_Y(u))$, we can rewrite this equation as

$$R_X(K_X(V(u))) = \sum_{ijk}\left(r_{ij}\kappa_{ki}\pi_i\pi_k u_k\right)\pi_j\phi_j = \sum_j w_j\pi_j\phi_j = V(w).$$

As a consequence,

$$R_X(K_X(V(u))) = V(R_Y(K_Y(u))).$$

Thus, as we have determined that $R_Y(K_Y(\tilde{m})) = \tau\tilde{m}$, it follows that

$$R_X(K_X(V(\tilde{m}))) = \tau V(\tilde{m}),$$

implying that $V(\tilde{m})$ is an eigenfunction of $R_X(K_X)$.

- From the definition of V, this eigenfunction can be written as

$$V(\tilde{m}) = \sum_i \pi_i^2 \nu_i \phi_i = \sum_i \frac{\nu_i}{\lambda_i} \phi_i.$$

□

The eigenfunction identified in Theorem 1 has the same form as the optimal discriminant operator introduced in Lemma 1, confirming the preservation of the properties of the Kurtosis matrix for the multivariate case, see [40], in the functional setting.

4 Computational Results

In this Section we present several results for the application of the proposed kurtosis operator to functional data, with the main goal of identifying clusters in the data. We have conducted simulation experiments, and we have also used publicly available data such as the Canadian Weather data set.
The implementation of our method has been carried out based on the **R** package *adf* which includes some utilities for Functional Data Analysis. The implementation has been conducted as described in Section 2.3, using both B-splines and Fourier functional bases.

4.1 Canadian Daily Weather

The *adf* package for **R** includes the *CanadianWeather* data set [53], consisting of daily measurements at 35 Canadian weather stations. The 35 Canadian weather stations are divided into four climate zones. In this example we have compared our classification results to these four distinct classes specified in the database: Atlantic, Pacific, Continental and Arctic.
The observation locations and the climate regions are located on the map of Canada shown in Figure 1, where the black diamonds correspond to the Arctic zone stations, the red color to Atlantic stations, the green color to Continental stations and the blue color to Pacific stations.

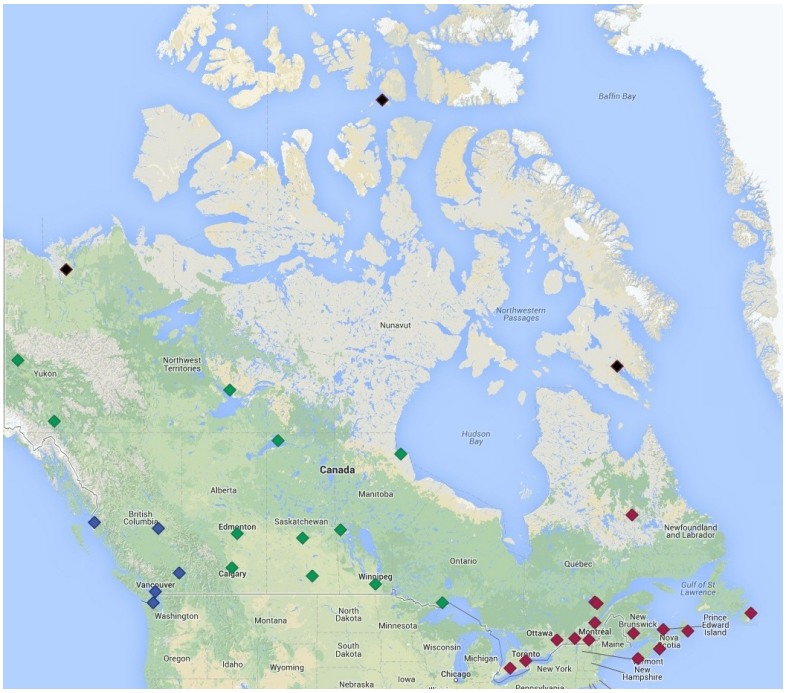

Figure 1: Canadian weather regions

We have used B-Spline and Fourier bases to represent the data, and after applying our procedure to estimate the kurtosis operator eigenfunctions, we have projected the data onto the two directions of maximum and minimum kurtosis, as well as those associated to the two largest (functional) principal components. The results are shown in Figures 2 and 3.

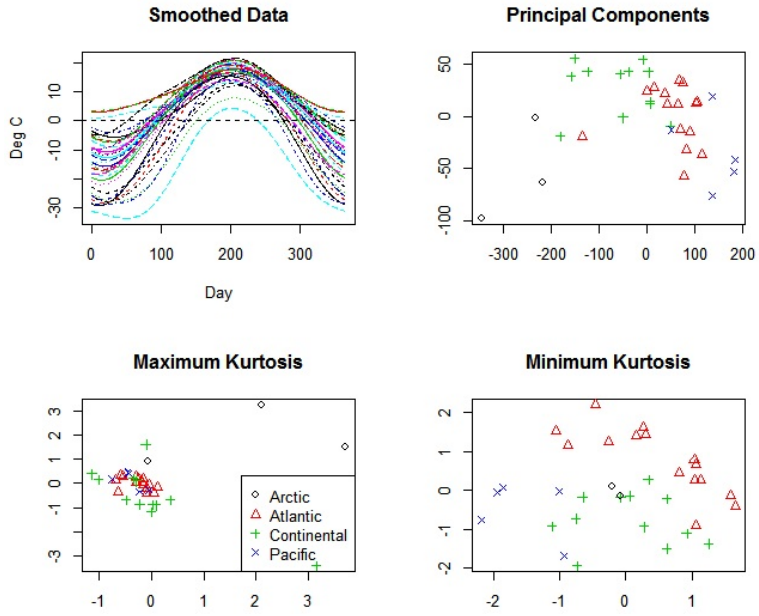

Figure 2: Fourier basis

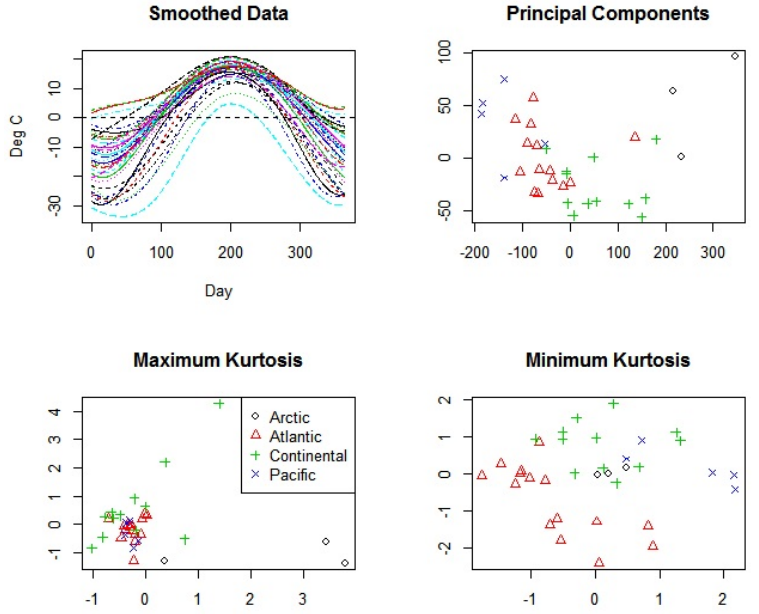

Figure 3: B-Spline bases

Our results provide a much better (although not perfect) separation between the observations corresponding to different regions, when compared to the groupings that could

be obtained from the principal components directions. In particular, the Atlantic, Continental and Pacific regions are clearly separated by the minimum kurtosis directions, while the Arctic region can be (at least partially) separated using the maximum kurtosis directions.

4.2 Simulated Data (Gaussian Processes)

We have performed two sets of simulations with the aim of comparing the performance of our proposed kurtosis operator with Functional Principal Components for unsupervised functional data clustering. The comparisons have been carried out using different versions of mixtures of gaussian processes. These models have been selected as they are simple ones and would allow us to verify a good fit to the theoretical results in Section 3. The analysis of the results should provide us with interesting insights on the behavior of the proposed method in a controlled environment.
In both cases the two populations of gaussian processes share the same quadratic covariance operator, $(\exp(-(x-y)^2/2l^2))$, with parameter $l = 15$. The same numbers of observations have been generated from each group ($n_1 = n_2 = n/2$). The observations have been obtained for $t \in [1, T]$ with $T = 20$. 20 equidistant observations of each process in $[1, 20]$ have been selected, with observation noise $\epsilon_{it} \sim N(0, 0.1)$. The values obtained are multivariate vectors in $\mathbb{R}^{20}$.
Both simulation examples differ in the choice of mean functions for each group, and in the processing of the information before applying our proposed procedure.
These generated data are then represented in the desired basis. We have used both Fourier and B-spline bases. From the smoothed data we have obtained the directions corresponding to the two largest eigenfunctions for the Functional Principal Components. We have also obtained the two directions corresponding to the smallest eigenfunctions of the kurtosis operator. We have projected our data onto these two pairs of directions.
To analyze the results, we have measured inter- vs. intra-group variability in the projections for each of the two groups, by comparing the traces of the corresponding covariance matrices. We have also applied K-means to the projected data and we have checked the classification results. Finally, we have prepared a graphical representation for one example of the clusters obtained by using principal component and kurtosis directions, to show how the kurtosis directions may be more efficient for cluster identification.
The basis used to represent data (Fourier or B-spline), the number of basis functions used and the number of observations for each group are modified between experiments. Each simulation experiment has been replicated 1000 times.

4.2.1 Simulation 1

In the first set of simulations (Simulation 1) we have used as mean functions for the two groups $m_i(t) = \sin(2\pi\mu_i t/T)$, $i = 1, 2$. The values μ_i are selected as -2.2 and 2, respectively.
For this example we wish to test if our method behaves reasonably well when the variability information has been removed from the data. To do that, and before fitting the

data to our chosen bases, we have introduced a linear transformation on the multivariate data so that the mean of the transformed sample is equal to zero and its covariance matrix is the identity. We expect functional principal components to have some difficulty separating the two modified groups. But note that principal components will work on the functional representation of the data, and may still capture some of that variability information. Our main interest is to check that kurtosis is able to identify the groups by using information beyond that of the variability in the data available through the covariance matrix.

Simulation 1. Fourier Basis Using a Fourier basis and the initial values mentioned above, we obtain the following results for inter- vs. intra-group variability shown in Table 1.

n	# **Bases**	**Variability Kurtosis**	**Variability PC**
30	7	0,64	0,01
60	7	0,78	0,06
180	7	0,90	0,06
30	15	0,22	0,01
60	15	0,36	0,01
180	15	0,64	0,01

Table 1: Fourier basis. Variability

Table 2 presents the proportion of misclassified observations using K-means. We have included a column ("Smoothed Data") corresponding to the application of K-means to the original smoothed data. That is, we have used in that column the multivariate data obtained from the functional representation of the data, observed at the initial data points. These results provide a reference for the advantages of using a functional representation of the data, as opposed to working directly with the data as multivariate observations.

n	# **Bases**	**Kurtosis Directions**	**PC. Directions**	**Smoothed Data**
30	7	0,15	0,46	0,49
60	7	0,16	0,42	0,27
180	7	0,13	0,42	0,22
30	15	0,34	0,46	0,39
60	15	0,28	0,46	0,38
180	15	0,15	0,47	0,34

Table 2: Fourier basis. K-means

Figures 4 and 5 show the plots corresponding to $n = 30$ and $n = 180$ respectively. The

projections have been obtained for the directions that minimize the kurtosis and principal components, using 7 functions in the basis representation.

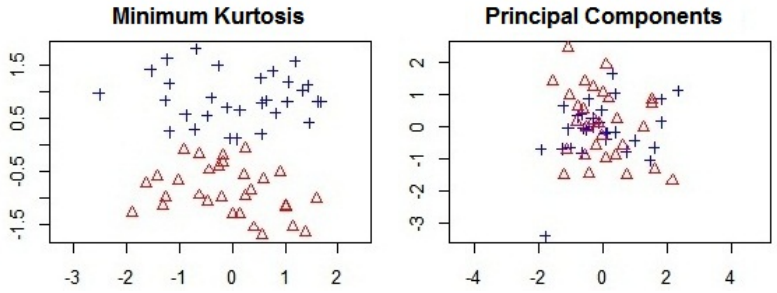

Figure 4: 7 Fourier basis. $n = 30$

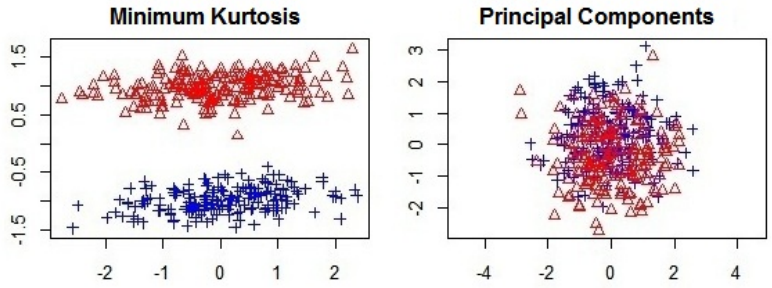

Figure 5: 7 Fourier basis. $n = 180$

The results obtained from the kurtosis directions are much better than those obtained from the principal components directions. It is also interesting to observe that the results for the kurtosis directions worsen with an increase in the basis size. We believe this is a direct consequence of the behavior of the kurtosis procedure in the multivariate case, when the dimension of the data increases. Also, the results are much better than those for the smoothed data, implying a clear advantage of the use of functional representations for the data.

Simulation 1. B-Spline Bases For the next set of results we use a B-Splines basis and the same values as in the preceding experiment. We obtain the results for inter- vs. intra-group variability shown in Table 3.

n	# Bases	Variability Kurtosis	Variability PC
30	7	0,50	0,46
60	7	0,62	0,46
180	7	0,75	0,47
30	15	0,22	0,01
60	15	0,36	0,01
180	15	0,64	0,01

Table 3: B-Spline bases. Variability

The proportion of misclassified observations using K-means, including those for the smoothed data, are given in Table 4.

n	# Bases	Kurtosis Directions	PC. Directions	Smoothed Data
30	7	0,24	0,23	0,32
60	7	0,19	0,22	0,32
180	7	0,13	0,18	0,32
30	15	0,34	0,46	0,47
60	15	0,28	0,46	0,48
180	15	0,15	0,46	0,49

Table 4: B-Spline bases. K-means

Figures 6 and 7 show the plots corresponding to $n = 30$ and $n = 180$ respectively. The projections have been obtained for the directions that minimize the kurtosis and principal components, using 7 functions in the basis representation..

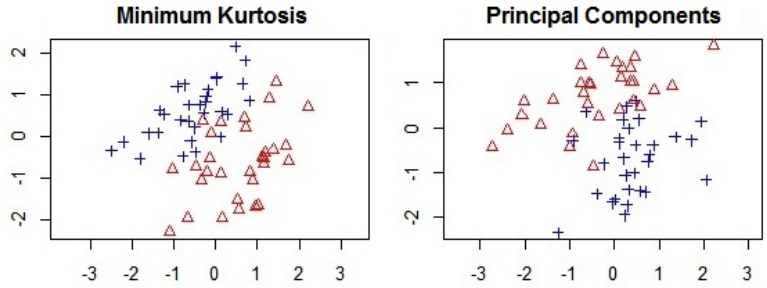

Figure 6: 7 B-Spline bases. $n = 30$

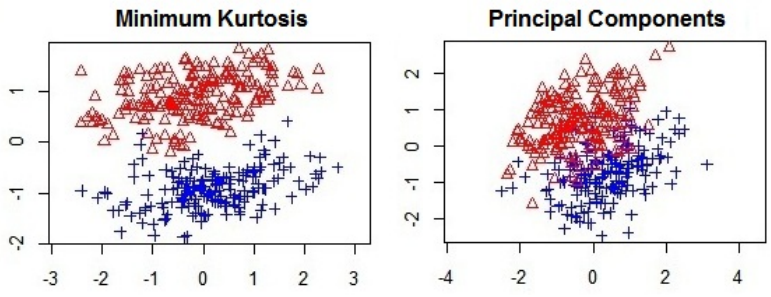

Figure 7: 7 B-Spline bases. $n = 180$

The results obtained for the proposed kurtosis method using a B-spline basis are interestingly worse than those using a Fourier basis. We believe this may be due to the basis providing a worse representation for the objects of interest (mean functions, covariance operator). This implies that the performance of the kurtosis operator may be sensitive to the choice of basis, at least in some cases, although this dependence would require a more detailed analysis (see the results for Simulation 2).

4.2.2 Simulation 2

We conduct a second experiment, similar to the preceding one, where we use mean functions equal to zero for the first group, and $0.2\cos(2\pi t/(T/r))$, for $r = 1.5$. Again, we have used a Fourier basis representation and a B-Splines basis; in both simulations the number of functions chosen for the basis is equal to 7. We have not included other basis sizes, as the preceding experiment seemed to indicate that this was a reasonable choice. In this case we have not carried out any additional transformation of the multivariate data. Our goal is now to test how well our proposed method performs when compared with functional principal components, if variability information is available in the covariance matrix to help classify the data. In this case we still expect our method to perform reasonably well, as we are using a model under which we have shown the proposed method has good classification properties. We wish to compare how much difference there may be between the use of the functional principal component directions and the kurtosis directions to reveal heterogeneity in the data.

Simulation 2. Fourier Basis Using a Fourier basis and the values mentioned above, we obtain the results for inter- vs. intra-group variability shown in Table 5.

n	**Variability Kurtosis**	**Variability PC**
30	0,46	0,12
60	0,56	0,11
180	0,68	0,11

Table 5: Fourier basis. Variability

In Table 6 we present the proportion of misclassified observations, including the results from the smoothed data, using K-means.

n	**Kurtosis Directions**	**PC. Directions**	**Smoothed Data**
30	0,25	0,47	0,47
60	0,21	0,48	0,48
180	0,13	0,49	0,48

Table 6: Fourier basis. K-means

Figures 8 and 9 show the graphs corresponding to $n = 30$ and $n = 180$ respectively, for the directions that minimize the kurtosis and principal components.

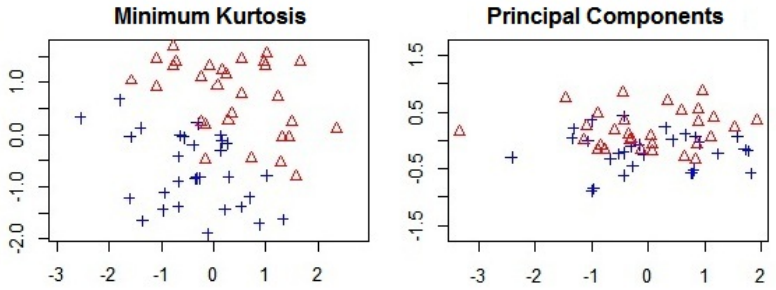

Figure 8: 7 Fourier basis. $n = 30$

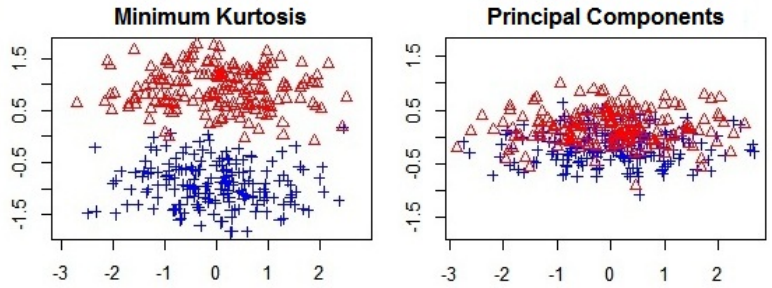

Figure 9: 7 Fourier basis. $n = 180$

If we compare these results with those from Simulation 1, we can see that we again obtain significantly improved results with respect to principal components. Also, and somewhat surprisingly, that functional principal components does not work much better than using multivariate techniques on the smoothed data. Finally, it seems interesting to note that the performance of the proposed method increases markedly with the sample size.

Simulation 2. B-Spline Bases Using a B-Splines basis and the values mentioned above, we obtain the results for inter- vs. intra-group variability shown in Table 7.

n	**Variability Kurtosis**	**Variability PC**
30	0,49	0,02
60	0,61	0,01
180	0,73	0,01

Table 7: B-Spline bases. Variability

In Table 8 we present the proportion of misclassified observations using K-means, including its application to the smoothed data.

n	**Kurtosis Directions**	**PC. Directions**	**Smoothed Data**
30	0,24	0,47	0,44
60	0,18	0,48	0,46
180	0,14	0,49	0,47

Table 8: B-Spline bases. K-means

Figures 10 and 11 show the graphs corresponding to $n = 30$ and $n = 180$ respectively, for the directions that minimize the kurtosis and principal components.

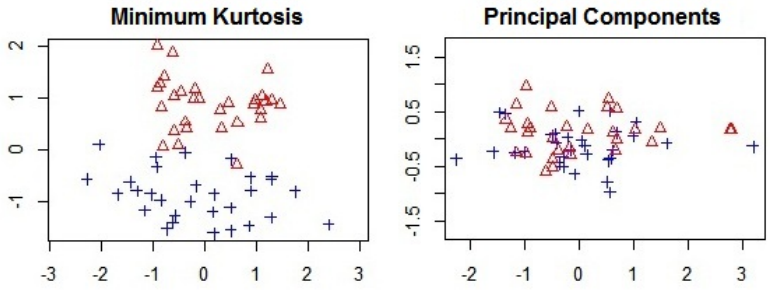

Figure 10: 7 B-Spline bases. $n = 30$

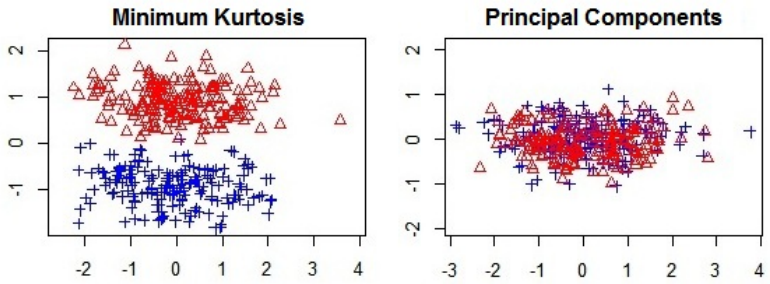

Figure 11: 7 B-Spline bases. $n = 180$

In this case, the dependence of the results on the choice of basis is very small, as the values we obtain are nearly identical with both basis choices. If anything, they seem to be even slightly better for the B-splines basis.
In summary, from the results in the preceding tables it follows that, under the models considered in the experiments, using the kurtosis directions provides an efficient way to reduce the dimension in the data without affecting its heterogeneity. It also provides a powerful tool for the exploratory analysis of these data.
These properties offer a marked improvement on the equivalent results obtained using functional principal components, or analyzing the data directly as multivariate observations. Thus, we believe that at least in some cases our proposed method provides clear advantages for the study of heterogeneous data, and the application of unsupervised classification techniques to these data.

5 Conclusions

In this paper we have introduced a kurtosis operator for functional data, based on the multivariate kurtosis matrix proposed by Móri et al [34]. We have also indicated the manner in which we implement the method and we have compared it with a multivariate alternative on the original data.
The theoretical properties of the kurtosis operator with respect to the classification of gaussian processes are very good (optimal in the case of gaussian processes with the same covariance operator) and inherit the corresponding properties of the multivariate proposal studied in [40].
In our simulation experiments we have shown that the proposed operator is able to outperform the behavior of the functional principal components operator regarding unsupervised classification, at least in some cases.
In summary, the proposed method is an interesting contribution to complement the information that can be extracted by applying more conventional methods, such as functional principal components, both to identify structures removed from normality (as in other Independent Component Methods) and in particular to identify clusters that might appear to be masked with respect to their variabilities.

References

[1] Abraham, C., Cornillon, P. A., Matzner-Lüber, E. and Molinari, N. (2003): *Unsupervised curve clustering using B-splines.* Scandinavian Journal of Statistics, 30, p. 581-595.

[2] Ali, M. M. (1974): *Stochastic ordering and kurtosis measure.* Journal of the American Statistical Association, 69, p. 543-545.

[3] Baíllo, A., Cuesta-Alberto, J. and Cuevas, A. (2011): *Supervised Classification for a Family of Gaussian Functional Models.* Scandinavian Journal of Statistics, 38, p. 480-498.

[4] Balanda, K.P. and MacGillivray, H.L. (1988): *Kurtosis: A critical review.* The American Statistician, 42, p. 111-119.

[5] Biau, G., Bunea, F. and Wegkamp, M. (2005): *Functional classification in Hilbert spaces.* IEEE Transactions on Information Theory. 51, p. 2162-2172.

[6] Cardoso, J. F. (1989). *Source separation using higher order moments.* Proc. ICASSP, p. 2109-2112.

[7] Chiou, J.M. and Li, P.L. (2007): *Functional clustering and identifying substructures of longitudinal data.* Journal of the Royal Statistical Society, serie B, 69, p. 679-699.

[8] Chissom, B. S. (1970): *Interpretation of the Kurtosis Statistic.* The American Statistician, 24-4, p. 19-22.

[9] Darlington, R. B. (1970): *Is kurtosis really "Peakedness"?.* The American Statistician, 24-2, p. 19-22.

[10] DeCarlo, L. T. (1997): *On the meaning and use of kurtosis.* Psychological Methods, 2, p. 292-307.

[11] Díaz, C., García, P., Alonso, J., Martínez, J. and Taboada, J. (2012): *Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain).* Science of the Total Environment, 439, p. 54-61.

[12] Febrero M., Galeano, P. and González-Manteiga, W. (2007): *A functional analysis of NOx levels: location and scale estimation and outlier detection.* Computational Statistics, 22, p. 411-427.

[13] Febrero M., Galeano, P. and González-Manteiga, W. (2008): *Outlier detection in functional data by depth measures with application to identify abnormal NOx levels.* Environmetrics, 19, p. 331-345.

[14] Ferraty, F., and Vieu, P. (2003): *Curves Discrimination: A Nonparametric Functional Approach.* Computational Statistics and Data Analysis, 44, p. 161-173.

[15] Ferraty, F., and Vieu, P. (2006): *Nonparametric Functional Data Analysis: Theory and Practice.* Springer.

[16] Finucan, H. M. (1964): *A Note on Kurtosis.* Journal of the Royal Statistical Society, Ser. B, 26-1 p. 111-112.

[17] Galeano, P. (2010): *Cluster identification for functional data by projection pursuit.* ERCIM

[18] Hall, P., Poskitt, D. and Presnell, B. (2001): *A functional data-analytic approach to signal discrimination.* Technometrics, 43, p. 1-9.

[19] Hampel, F. R. (1974): *The influence curve and its role in robust estimation.* Journal of the American Statistical Association, 69, p. 383-393.

[20] Hartigan, J. A. and Wong, M. A. (1979): *A k-means clustering algorithm.* Journal of the Royal Statistical Society, Series C (Applied Statistics), 28, p. 100-108.

[21] Hyvärinen, A., Karhunen, J. and Oja E. M. (2001): *Independent Component Analysis.* New York: John Wiley.

[22] Hastie, T., Buja, A. and Tibshirani, R. (1995): *Penalized discriminant analysis.* The Annals of Statistics, 23, p. 73-102.

[23] Jacques, J. and Preda, C. (2012): *Model-based clustering of functional data.* In: 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, p. 459-464.

[24] James, G. M. and Sugar, C. A. (2003): *Clustering for sparsely sampled functional data.* Journal of the American Statistical Association, 98, p. 397-408.

[25] Kaplansky, I. (1945): *A common error concerning kurtosis.* Journal of the American Statistical Association, 40, p. 259.

[26] Liu, J., Zhang, J., Palumbo, M. and Lawrence C. (2003): *Bayesian clustering with variable and transformation selections.* Bayesian Statistics, 7, eds, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: University Press, p. 249-75.

[27] López-Pintado, S., and Romo, J. (2006): *Depth-based classification for functional data.* Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. American Mathematical Society. DIMACS Series, 72, p. 103-121.

[28] MacQueen, J. B. (1967): *Some methods for classification and analysis of multivariate observations.* Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, p. 281?297.

[29] Mardia, K. V. (1970): *Measures of multivariate skewness and kurtosis with applications.* Biometrika 57, p. 519-530.

[30] Martínez, J., Garcia P., Alejano, L. and Reyes, A. (2011): *Detection of outliers in gas emissions from urban areas using functional data analysis.* Journal of Hazardous Materials, 186, p. 144-149.

[31] Mercer, J. (1909): *Functions of positive and negative type and their connection with the theory of integral equations.* Philosophical Transactions of the Royal Society, 209, p. 415-446.

[32] Moors, J. J. A. (1986): *The Meaning of Kurtosis: Darlington Reexamined.* The American Statistician, 40, p. 283-184.

[33] Moors, J. J. A. (1988): *A Quantile Alternative for Kurtosis.* The Statistician, 37, p. 25-32.

[34] Móri, T. F., V. K. Rohatgi, and G. J. Székely (1993): *On multivariate skewness and kurtosis.* Theory of Probability and its Applications, 38, p. 547-551.

[35] Oja, H. (1981): *On Location, Scale, Skewness and Kurtosis of Univariate Distributions.* Scandinavian Journal of Statistics, 8-3, p. 154-168.

[36] Pearson, K. (1905): *Das Fehlergesetz und Seine Verallgemeinerungen Durch Fechner und Pearson. A Rejoinder.* Biometrika, 4, p. 169-212.

[37] Peña, D. (2002): *Análisis de datos multivariantes.* McGraw-Hill.

[38] Peña, D. and F. J. Prieto (2001): *Cluster identication using projections.* Journal of the American Statistical Association, 96, p. 1433 - 1445.

[39] Peña, D. and F. J. Prieto (2001): *Robust covariance matrix estimation and multivariate outlier detection (with discussion).* Technometrics, 43, p. 286-310.

[40] Peña, D., Prieto, F. J. and Viladomat, J. (2010): *Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure.* Journal of Multivariate Analysis, 101, p. 1995-2007.

[41] Pérez, C. and Santín, D. (2007): *Minería de Datos. Técnicas y Herramientas.* Thomson.

[42] Ramsay, J. O. and Dalzell, C. (1991): *Some tools for functional data analysis.* Journal of the Royal Statistical Society, Series B, 53, p. 539-572.

[43] Ramsay, J. O., Hooker, G., and Graves, S. (2009): *Functional Data Analysis with R and MATLAB (Use R).* Springer.

[44] Ramsay, J. O, and Silverman, B.W. (2002): *Applied Functional Data Analysis: Methods and case studies.* Springer.

[45] Ramsay, J. O, and Silverman, B.W. (1997): *Functional data analysis.* Springer.

[46] Ramsay, J. O, and Silverman, B.W. (2005): *Functional Data Analysis, Second Edition.* Springer.

[47] Ruppert, D. (1987): *What is kurtosis? An influence function approach.* The American Statistician, 41, p. 1-5.

[48] Serban, N. and Wasserman, L. (2004): *CATS: clustering after transformation and smoothing.* Journal of the American Statistical Association, 100, p. 990-999.

[49] Song, J., Lee, H., Morris, J. and Kang S. (2007): *Clustering of time-course gene expression data using functional data analysis.* Computational Biology and Chemistry, 31 (4), p. 265-274.

[50] Tyler, D. E., F. Critchley, L. Dumbgen, and H. Oja (2009): *Invariate co-ordinate selection (with discussion)*. Journal of the Royal Statistical Society: Serie B (Statistical Methodology), 71 (3), p. 1-27.

[51] Viladomat, J. (2012): *Illustrating a clustering algorithm based on a kurtosis matrix using stock market data*. MAF 2012.

[52] Yao, F., Müller, H.G. and Wang J.L. (2004): *Functional data dnalysis for sparse longitudinal data*. Journal of the American Statistical Association, 100 (470), p. 577-590.

[53] http://www.functionaldata.org/.